Clim. Past Discuss.,
doi:10.5194/cp-2016-129-AC1, 2017

Climate of the Past Discussions

Interactive comment on "Reconstructing past climate by using proxy data and a linear climate model" *by* Walter A. Perkins and Gregory J. Hakim

Walter A. Perkins and Gregory J. Hakim

wperkins@uw.edu

Received and published: 1 March 2017

1 Major Comments

I find that this article does not belong to the journal of Climate of the Past. Thus, I reject the publication. However, I find the idea novel and interesting and I would suggest to submit a revised version to a more theoretical journal. Here is why it doesn't belong to Climate of the Past: The idea of using LIM as a substitute for otherwise expensive online data assimilation sounds wonderful, and would prove to be it if one didn't need to introduce a parameter a. But as authors showed with a being equal to 1, the results of linear online DA are worse than of offline DA. Then the question I pose to authors is how to choose an optimal a?

What would be the criteria for optimal a?

We thank the referee for commenting on this paper. On this first point, we disagree with the assessment that this paper does not belong in the journal of Climate of the Past (CP), and our reply to that can be found below. On the need for a tuning parameter (such as a) in the usage of ensemble data assimilation methods, it is common due to the affect such parameters have on ensemble variance. As in Hamill and Snyder (2001), we present results for a range of the blending coefficient (a), but show that there is a general range of blending that gives improved validation metrics compared to the offline case. We do not give an absolute determination of an optimal (a) because, as we show in the paper, the choice of metric for determining the best value for (a) is subjective. The use of correlation, the coefficient of efficiency, and the continuous ranked probability score (CRPS) target different aspects of comparison between the reconstructed and reference data (though CE and CRPS are very similar). The details of the differences in different skill metrics are described in Section 3.1. Altogether, the determination of which metric is best depends on what the end user prioritizes.

There is yet another manifestation of more theoretical work to be done, mainly in Sec. 4.2, where authors find inconsistency between the best model 20CR in terms of a scalar skill (CE, r, and CRPS) but worse in terms of spatial reconstruction. They propose that it could be due to a short time scale but this could be checked. And again, does this mean that $a = 0.9$ is not optimal for 20CR?

We agree that this result can be more thoroughly discussed. The 20CR dataset is a reanalysis performed using surface pressure observations from 1850 to present (Compo et al. 2011). The observational coverage of the southern oceans, especially during the early portion of the record is very low (see Fig. 3 in Woodruff et al. 2010). This is a large region of high variability due to the southern hemisphere storm track. As Fig. 7 in our paper shows, the primary mode of spatial variability in the annually averaged 20CR dataset is quite different in character from the other 3 datasets used. The pattern of variability is focused on regions in the southern oceans, which is where we see

a lot of the CE skill degradation. This coupled with the knowledge that it is a region of high variability with few observations leads us to speculate that the features of the 20CR LIM may be influenced by artifacts of the 20CR dataset. Another instrumental reanalysis product spanning 1900 – 2012 (ERA-20C; Poli et al. 2016) has a similar first EOF and forecast mode to the 20CR. The discrepancy between the GMT and spatial performance for the 20CR experiment was surprising, but similar results were found in idealized pseudo-proxy experiments (Annan Hargreaves 2012, Wang et al., 2014). These studies show that spatial averaging can boost the signal-to-noise ratio for large-scale indices resulting in higher index skill despite poor spatial reconstruction performance. Here we have a situation that is slightly different. The spatial results for the offline case are decent, but the degradation of spatial skill by the 20CR LIM reconstruction enhances the skill in the GMT signal. We interpret this as the spatial degradation having a moderating effect on an aspect of the global signal that is overestimated in the offline case such as the interannual variance or the warming trend. We can provide a breakdown of this idea with plots of calculated spatial trends and interannual GMT variance over time. Again, $a = 0.9$ is the optimal blending when considering CE skill of GMT. Changing the blending parameter will not change the behavior of the forecasts, only how much information is used from the forecast. If a user was inclined to use the 20CR LIM for a reconstruction, they could optimize between spatial skill and GMT skill, but the use of 20CR LIM forecast information will only reduce spatial skill. From what we show in the paper, there are better options than the 20CR LIM to use for spatial reconstruction performance.

Therefore, what I suggest is to study the methodology in a theoretical framework by revising the article and submitting it to a theoretical journal.

Regarding the larger question of relevance to this journal, we believe that there exists well-established precedent for studies like the current one. The field of paleoclimate data assimilation (PDA) has had a number of recent theoretical advances, many of which were published in Climate of the Past (CP). For example, Crespin et al. (2009)

extend the ensemble selection DA method to perform online forecasts with an intermediate complexity climate model. However, they make no comparison with its offline predecessor. Bhend et al. (2012) use an idealized pseudo-proxy experiment to test an offline ensemble square root filter approach. In this work, they use a parameter known as covariance localization to prevent the collapse of ensemble variance, and they also discuss the extension to a coupled model forecast online method as the next step. Annan and Hargreaves (2012) perform an idealized pseudo-proxy experiment with an ensemble selection method to test different geographic densities of proxy observations. They also try a persistence forecast method as an online case, but find no benefits to the reconstruction from the persistence forecast in all cases. Matsikaris et al. (2015) perform a direct comparison between online and offline methods using the ensemble selection method with a 10-member ensemble forecast from a coupled GCM. They also find no discernible benefits to using the online method compared to the offline experiment. They follow up with a study (Matsikaris et al., 2016) further investigating why the online forecast method does not show improvements to their reconstruction targets. Thus it is clear that there are numerous papers in CP similar to ours; this is why we submitted the paper to CP. We have presented the only study to our knowledge that shows a viable online PDA method that can be used for long timescales with large ensembles while also documenting improvements over the offline alternative. Additionally, we document the performance of the method using real proxy data in real reconstructions. We plan to provide code and sample data along with the revised manuscript. The connection to previous CP literature and our novel results make this study both timely and relevant for this journal.

LIM is calibrated on model 1 (CCSM4) without any data assimilation over a period 850-1850, on model 2 (MPI) without any data assimilation over a period 850-1850, on model 3 with data assimilation (20CR) over a different period 1850-2012, and on a data set over yet a different time period (1950-2010 I would assume, though it is not mentioned in the paper). Thus, the models are completely different in terms of the time period, use or not of the data, and only being the data. This

makes it hard to compare and draw conclusions. Instead LIM should be calibrated on a model without DA, on the same model with DA, and on observations used in that DA.

That would be a nice framework for a future study, but it is clearly well beyond the scope of this paper. In any case, our method has sampled widely different sources for variability on which to calibrate the LIMs. During the instrumental period the separation between forced and natural variability is unclear. However, the last millennium simulations are a good substitute for a measure of long-term natural variability under weaker forcing regimes. As far as the use or not of observations in the calibration data, we will note that for climate models there are no common systems that provide data assimilation for simulations like this. That means a comparison like the one suggested would have to be done using systems oriented for reanalysis, which are very different from climate models. Our results show that reanalysis products may have inherent properties making them less useful for our application. Additionally, the observations used for assimilation during the instrumental period are not necessarily something that we can use as a LIM forecast model in reconstruction. For example, the 20CR assimilates pressure observations. Other reanalysis products use a suite of different measurements including satellite radiances and upper air measurements. These are not easily useable gridded products and are outside the scope of variables we would use in paleoclimate applications. As we said, these are topics that can be explored in future research. We will change the text to explicitly define that the BE dataset covers the years 1960-2014.

As the prior authors used results of the CCSM4 model, the same model they used for LIM calibration. It appears that linear CCSM4 DA provides good results in terms of both scalar skills and spatial reconstruction. Is it because there is less inconsistency? How would it change if the prior was from another model?

We did check results using the MPI last millennium simulation as our prior and the NOAA merged land-ocean surface temperature analysis (MLOST) as the calibration

data for the proxy observation models. The CCSM4 LIM still outperformed the MPI LIM by a similar margin in the spatial results (the CCSM4 spatial average CE was +.02 above the MPI case). The GMT skill results show the two models again at a virtual tie. This suggests that the CCSM4 LIM provides forecasts that generate results more consistent with the GISTEMP reference we use for validation.

In order to provide a fair comparison authors need to include "expensive" online DA (using a nonlinear model instead of LIM).

As stated in our introduction, the current computational costs of performing coupled model forecasts in large ensembles over long time periods make this suggestion impractical. Moreover, other studies have explored this possibility (with smaller ensembles forecasting on decadal timescales) and shown no benefit over offline DA. This aspect and the knowledge that a LIM can be comparable in forecast skill to the coupled models (Newman 2013) is why we chose to try a LIM for online paleoclimate data assimilation. We believe this is a fair baseline comparison showing the potential of an online method compared to an offline method. Despite the simplicity of this approach, we show improvements in reconstruction skill over the offline method, which is the first to our knowledge for any online technique. "Expensive" online PDA options likely won't be feasible with ensembles of the size we use here anytime soon.

2 Minor Comments

Page 7, Line 18: Why is there a shift in blending coefficient? This is again related to my comment on how to choose an optimal a.

There is a shift in blending coefficient between CE and CRPS because they are different skill metrics. CE is calculated based on mean squared error properties, while CRPS is calculated on accumulated mean absolute error. Though they are both sensitive to bias, phase, trend, and amplitude differences, there are no guarantees that they

will give the same results. This is especially apparent for detrended GMT skill of the persistence case when $a = 1.0$.

Page 7, Line 31: Why is there improvement compared to offline DA even though the trend is largely underestimated for α = 0.95?

There is improvement because the trend is not the only consideration of the CRPS/CE skill metrics. Other aspects (phase and amplitude) can still be improving while the trend difference is not large enough to decrease the skill measure.

It would be interesting to introduce another metric – bias – in order to check whether the model either underestimates or overestimates the observed values.

The CE metric can be separated to show skill related specifically to bias. We will consider expanding the GMT and spatial breakdown of CE skill between bias and other aspects of the reconstruction.

I suggest plotting time series of averaged temperature of different models against observations for best a for CE, for best a for r, and for best a for CRPS.

We agree that we should include a figure of the actual reconstructed GMT for selected reconstructions. Thank you for this suggestion.

3 References

Annan, J. D., Hargreaves, J. C. (2012). Identification of climatic state with limited proxy data. Climate of the Past, 8(4), 1141–1151. http://doi.org/10.5194/cp-8-1141-2012

Bhend, J., Franke, J., Folini, D., Wild, M., Brönnimann, S. (2012). An ensemble-based approach to climate reconstructions. Climate of the Past, 8(3), 963–976. http://doi.org/10.5194/cp-8-963-2012

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., ... Worley, S. J. (2011). The Twentieth Century Reanalysis Project. Quarterly Journal of the Royal Meteorological Society, 137(654), 1–28. http://doi.org/10.1002/qj.776

Crespin, E., Goosse, H., Fichefet, T., Mann, M. E. (2009). The 15th century Arctic warming in coupled model simulations with data assimilation. Climate of the Past, 5(3), 389–401. http://doi.org/10.5194/cp-5-389-2009

Hamill, T. M., Snyder, C. (2000). A Hybrid Ensemble Kalman Filter–3D Variational Analysis Scheme. Monthly Weather Review, 128(8), 2905–2919. http://doi.org/10.1175/1520-0493(2000)128<2905:AHEKFV>2.0.CO;2

Matsikaris, A., Widmann, M., Jungclaus, J. (2015). On-line and off-line data assimilation in palaeoclimatology: a case study. Climate of the Past, 11(1), 81–93. http://doi.org/10.5194/cp-11-81-2015

Matsikaris, A., Widmann, M., Jungclaus, J. (2016). Influence of proxy data uncertainty on data assimilation for the past climate. Climate of the Past, 12(7), 1555–1563. http://doi.org/10.5194/cp-12-1555-2016

Newman, M. (2013). An Empirical Benchmark for Decadal Forecasts of Global Surface Temperature Anomalies. Journal of Climate, 26(14), 5260–5269. http://doi.org/10.1175/JCLI-D-12-00590.1

Poli, P., Hersbach, H., Dee, D. P., Berrisford, P., Simmons, A. J., Vitart, F., ... Fisher, M. (2016). ERA-20C: An atmospheric reanalysis of the twentieth century. Journal of Climate, 29(11), 4083–4097. http://doi.org/10.1175/JCLI-D-15-0556.1

Wang, J., Emile-Geay, J., Guillot, D., Smerdon, J. E., Rajaratnam, B. (2014). Evaluating climate field reconstruction techniques using improved emulations of real-world conditions. Climate of the Past, 10(1), 1–19. http://doi.org/10.5194/cp-10-1-2014

Interactive comment on Clim. Past Discuss., doi:10.5194/cp-2016-129, 2016.